FIS - Statistical Machine Learning
Assignment 2

Please send me

- the **original script** detailing your computations.
  - The script must be **documented**, i.e. the code corresponding to each answer must be delimited and your loops/variables briefly explained.
  - The script must be **executable**: by just running your script, all results should appear **automatically**.
  - You can use external functions such as optimization and statistics toolboxes. If you do so, name explicitly the toolboxes that you use. When in doubt ask me, but code things preferably **by yourself**.
- A **document** (.doc, .pdf) which will contain your answer and your analysis. Do not put your source code in that document. Illustrate your answers with graphs and plots.

This homework is due **July 16th (Sat.) 23:59 PM**

Send your homework to mcuturi@i.kyoto-u.ac.jp

Exercise 1: Classification - Hoeffding's and V.C Bounds

- Choose two gaussian densities p_{-1}, p_{+1} on $\mathbb{R}$ with unit variance and mean in $[-1, 1]$. We consider a pair of random variables (X, Y) where the density of (X, Y) is defined by the following: $p(Y = 1) = 0.65$ and the density of $p(X|Y = 1)$ is equal to p_{+1} while $p(X|Y = -1)$ is equal to p_{-1}.
- Consider $N = 20$ different linear classifiers on $\mathbb{R}$, that is step functions defined by a threshold τ and a sign $t \in \{-1, 1\}$ as

$$f_{t,\tau}(x) = \begin{cases} t \text{ if } x > \tau \\ -t \text{ if } x \leq \tau \end{cases} .$$

  Choose t randomly and select all thresholds τ in $[-2, 2]$.
- Give a detailed illustration of Hoeffding's bound for the supremum of the difference of the empirical risk and the true risk for the set of N functions considered above, by sampling 200 sets of $n = 20, 50, 100$ independent observations of (X, Y). In order to do so, you will need to compute the true risk of each of the Heaviside functions (the Error function might be useful) and sample randomly from the densities p_{-1} and p_{+1}. Try to split these steps using short subroutines.

- We have studied Vapnik Chervonenkis bounds for infinite families of functions. Give an expression for this bound when considering all possible translations and multiplications by $\{-1, 1\}$ of the Heaviside-functions. Your bound should only depend on the threshold ε and sample size n. Find a condition on N for which the VC bound is tighter (that is, provides a lower bound) than Hoeffding's bound.

Exercise 2: Support Vector Machines

- Download a binary classification data-set on the UCI Machine Learning Repository. Choose a dataset with a few attributes (less than 50 approximately) and only keep at most 1000 observations.

- If this has not been done, split the dataset into test and training folds of equal or similar size. The two folds should be **balanced** in the sense that they should contain (approximately) the same proportion of positive and negative labels than your original dataset.

- Implement the dual formulation of soft-margin SVM's using any quadratic program solver (quadprog in R, quadprog or CVX in Matlab, CVX-OPT in Python). Your code should take 3 inputs: Gram matrix K, vector of labels y and constant C.

- Consider the Gaussian kernel with parameter σ. Consider the following parameters:

  - 7 different values for σ: $\{.05, .1, .5, 1, 5, 10, 50\}\cdot\text{Median}(\|x-x'\|)$, where the median can be sampled on a subset of your training set,
  - 7 different values for C, $\{.001, .01, .1, 1, 10, 100, 1000\}$.

  Compute the train error for every 7×7 couple of σ and C parameters and display these results graphically, using a loglog contour plot or a colored matrix.

- For each of these 7×7 parameter settings, compute also the test error, compute the test error obtained after training the classifier on the training fold, and display results in the same way. Compare the train and test errors. Does the setting with the lowest train error coincide with that of the lowest test error?